Nakao *et al. Earth, Planets and Space (2023) 75:78*
https://doi.org/10.1186/s40623-023-01839-y

Earth, Planets and Space

EXPRESS LETTER

Open Access

Regression analysis and variable selection to determine the key subduction-zone parameters that determine the maximum earthquake magnitude

Atsushi Nakao[1,4*], Tatsu Kuwatani[1], Kenta Ueki[1], Kenta Yoshida[1], Taku Yutani[1], Hideitsu Hino[2] and Shotaro Akaho[2,3]

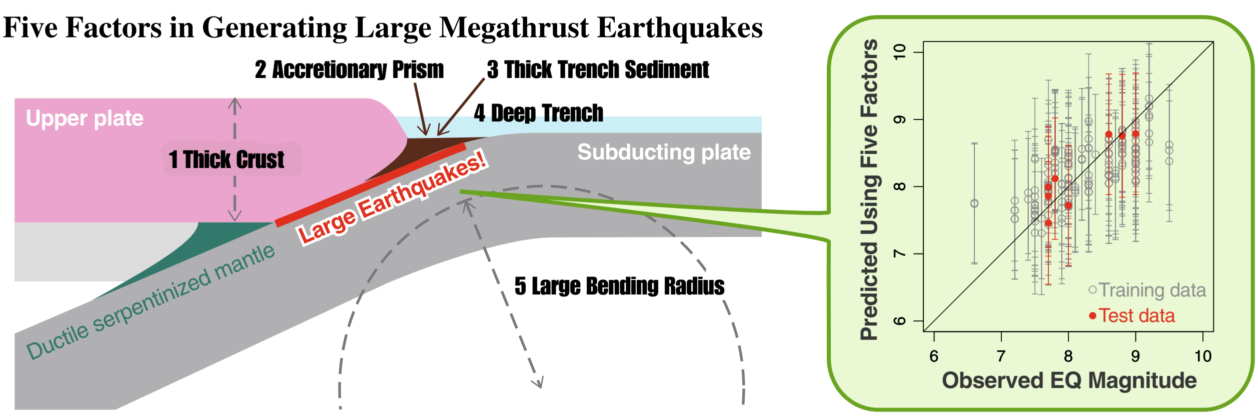
Abstract

Large variations in the maximum earthquake magnitude (M_{max}) have been observed among the world's subduction zones. There is still no universal relationship between M_{max} and a given subduction-zone parameter, such as plate age, plate dip angle, or plate velocity, which suggests that multiple parameters control M_{max}. Here, we conduct exhaustive variable selections that are based on three evaluation criteria; leave-one-out cross-validation errors (LOOCVE), Akaike information criterion (AIC), and Bayesian information criterion (BIC) to determine the combination of subduction-zone parameters that best explains M_{max}. Multiple linear regression analyses are applied using 18 subduction-zone parameters as potential candidates for the explanatory variables of M_{max}. The minimum BIC is obtained when five variables (trench sediment thickness, existence of an accretionary prism, upper-plate crustal thickness, bending radius of the subducting oceanic plate, and trench depth) are selected as explanatory variables; each variable contributes positively to M_{max}. Minimum LOOCVE and AIC values are obtained when eight variables (the five parameters for BIC, plus the along-strike plate convergence rate, age of the subducting plate, and maximum depth of the subducting plate) are selected. Our selection of the trench sediment thickness and plate bending radius contributing to M_{max} is consistent with previous studies. The results show that increasing upper-plate crustal thickness results in a large M_{max}. In addition to smoothing the subducting-plate interface via subducted sediments, along-dip extension of the crustal area along the convergent plate boundary would be important for generating a large earthquake.

Keywords Earthquake magnitude, Subduction zone, Multiple regression analysis, Exhaustive model evaluation, Plate tectonics

*Correspondence:
Atsushi Nakao
a-nakao@gipc.akita-u.ac.jp; atsushi.nakao@gmail.com
Full list of author information is available at the end of the article

Graphic Abstract



Introduction

Large earthquakes (magnitude $M \geq 8$) rarely occur, but are often detrimental to life and property. They have usually been observed along subduction zones and some continental-collision zones, with a significant variation in the maximum earthquake magnitude (M_{max}) detected among the world's subduction zones (Fig. 1); for example, the 1960 M9.5 Chile earthquake along the South-Central Chilean subduction zone is the largest recorded earthquake to date, whereas a M-7+ event has not been observed along the South Kermadec subduction zone. Seismologists generally assume that large earthquakes are associated with certain subduction settings, with numerous relationships between M_{max} (or the Gutenberg–Richter b-values) and various parameters that characterize the tectonic features of subduction zones (hereafter referred to as "subduction-zone parameters") proposed (e.g., Wirth et al. 2022; Marzocchi et al. 2016); for example, the age of the subducting plate (e.g. Ruff and Kanamori 1980; Nishikawa and Ide 2014), angle or curvature radius of the subducting plate (e.g., Ruff and Kanamori 1980; Bletery et al.

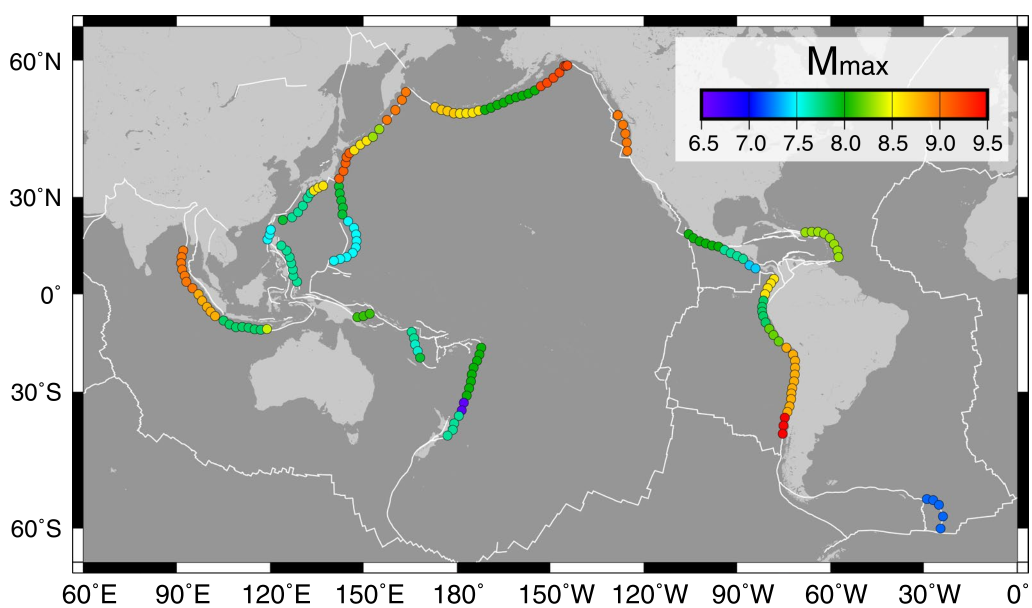


Fig. 1 Map of the observed maximum earthquake magnitude, M_{max} (unstandardized). The data is referred from SubMap 4.3 and includes the 169 locations. The white lines indicate plate boundaries

2016), seafloor sediment thickness at the subduction trench (e.g., Ruff 1989; Heuret et al. 2012; Scholl et al. 2015; Brizzi et al. 2018), subducted sediment thickness (Seno 2017), fore-arc structures (e.g., Song and Simons 2003; Wells et al. 2003), upper-plate strain (e.g., Heuret et al. 2012), trench migration velocity (e.g., Schellart and Rawlinson 2013), upper-plate motion (e.g., Scholz and Campos 1995), width of the subducting plate or trench length (Schellart and Rawlinson 2013; Brizzi et al. 2018), and topographic roughness or seafloor smoothness along the subducting plate (e.g., Wang and Bilek 2014; Lallemand et al. 2018) have all been analyzed to infer M_{max}. Schellart and Rawlinson (2013) have investigated 24 physical parameters that characterize subduction zones, but were unable to find any parameters that had a large correlation with M_{max} (correlation coefficient less than 0.5). Thus, there is still no consistent relationship between M_{max} and these individual subduction-zone parameters, which suggests that multiple factors may be involved. Alternatively, observational errors in subduction zone parameters may make it difficult to relate such parameters to M_{max}, and it is therefore necessary to select parameters with a low signal-to-noise ratio to correctly predict M_{max}.

The ability to determine the key subduction-zone parameters that influence the occurrence of large earthquakes is limited by the ability to effectively derive a small number of essential elements from a desired phenomenon. An exhaustive variable selection procedure combined with regression or discriminant analysis, which is a primitive machine-learning-based method, is a powerful approach to derive a small number of essential variables from complex processes (e.g., Kuwatani et al. 2014; Igarashi et al. 2018; Ueki et al. 2020; Itano et al. 2020; Nakao et al. 2022). This method is suitable for determining a combination of subduction-zone parameters that can reasonably explain M_{max}, and can be instrumental in gaining scientific insight into the origin of large earthquakes; however, this machine-learning approach has not been applied to derive a M_{max} relationship to date. Variable selection is a reasonable approach for addressing the present problem for two key reasons. First, the sample locations for subduction-zone parameters are limited. For example, the observed subduction styles, including the velocity and shape of the subducting plate, exhibit much smaller variations than those simulated in laboratory and numerical experiments (e.g., Schellart 2011; Nakao et al. 2016). Model selection with cross-validation would also be useful in enhancing the predictability of M_{max} with limited observations. Second, the observed subduction-zone parameters, including the velocity and geometry of the subducting oceanic plate, contain large uncertainties, such that a model may be overfit due to these uncertainties if a variable selection approach is not employed. Therefore, we employ an exhaustive variable selection approach in this study to infer which subduction-zone parameters may explain local variations in M_{max}.

Data and methods

Data

We investigate 18 types of subduction-zone parameters via regression analysis to determine the set of parameters that can effectively constrain M_{max}. Subduction-zone parameters are sampled at 2-degree intervals following Heuret and Lallemand (2005), at which both M_{max} and its explanatory variables are sampled. We incorporate present-day observations in our analysis, and therefore evaluate potential M_{max} values under present-day tectonic conditions.

The objective variable, M_{max}, is taken from the SubMap 4.3 database and is based on the rupture areas of large subduction earthquakes ($M \geq 8$) that occurred at depths less than 70 km during the 1900–2007 period (Heuret et al. 2011), as well as the 2011 Tohoku-oki Earthquake (M9.1, Northeast Japan; Yagi and Fukahata 2011). In addition, we included known historical earthquakes, including the 1700 Cascadia earthquake (M9.0, North America; Satake et al. 1996), the 1707 Hoei earthquake (M8.6, Southwest Japan; Fujiwara et al. 2020), and the 1833 Sumatra earthquake (M8.8, Indonesia; Zachariasen et al. 1999). The segmentation of M_{max} is defined based on three criteria of Heuret et al. (2011): (1) the rupture area inferred for M+8.0 earthquakes must be included in a single segment; (2) the transects with homogeneous activity in the seismogenic zone were grouped; and (3) the transects with homogeneous geometries in the seismogenic zone were grouped.

The subduction-zone parameters that we investigate in relation to M_{max} are listed in Table 1 and shown graphically in Fig. 2, with M_{max} values obtained at 169 locations along subduction zones worldwide (Fig. 1). We mainly employed the subduction-zone parameters from the SubMap 4.3 database (e.g., Heuret and Lallemand 2005) in our analysis. This data set lacks observations along some subduction zones (e.g., Mediterranean subduction zones); therefore, we excluded all sample locations with at least two missing variables. We conducted leave-one-out cross validation, a method to enhance the model predictability for unknown samples, to minimize the effects of this omission. We employed a neighboring value at the locations with only one missing variable. We consider that the effect to be small because the operation was applied to only 0.3% of the total data, and because the subduction-zone parameters generally vary continuously along a trench. Furthermore, we removed

Table 1 Analyzed subduction-zone parameters in this study

Symbol	Explanation	Unit	m_i [a]	s_i [b]	References
A	Age of subducting plate	Ma	67.28	42.29	Müller et al. (1997)
a_s	Dip angle of subducting plate	degree	30.21	10.75	Heuret and Lallemand (2005)
CMP	Dummy variable for compressive upper plate	–	0.2189	0.1720	Heuret et al. (2011)
MT	Dummy variable for accretionary prism	–	0.4260	0.4960	Brizzi et al. (2018)
R_c	Bending radius of subducting plate	km	407.2	198.4	Heuret (2005)
R_{IW}	Intermediate-wavelength seafloor roughness	m	586.9	428.0	Lallemand et al. (2018)
R_{LW}	Long-wavelength seafloor roughness	m	445.1	398.9	Lallemand et al. (2018)
R_{SW}	Short-wavelength seafloor roughness	m	139.5	70.83	Lallemand et al. (2018)
T_c	Upper-plate crustal thickness	km	31.12	16.24	Laske et al. (2013)
TNS	Dummy variable for tensile upper plate	–	0.2781	0.2020	Heuret et al. (2011)
T_{sed}	Trench sediment thickness	km	0.6568	0.7641	Straume et al. (2019)
v_{sn}	Convergence rate at trench (trench-normal)	mm/y	57.16	30.67	Lallemand et al. (2008)
v_{ss}	Convergence rate at trench (trench-parallel)	mm/y	21.30	16.80	Lallemand et al. (2008)
v_{dn}	Upper-plate extension rate	mm/y	− 5.148	25.00	Lallemand et al. (2008)
v_{tn}	Trench retreat rate	mm/y	8.189	28.10	Lallemand et al. (2008)
Z_{seis}	Maximum earthquake depth	km	359.4	207.3	Heuret (2005)
Z_t	Trench depth	km	5.787	1.787	Heuret (2005)
Z_{tomo}	Maximum slab depth	km	639.9	308.7	Heuret and Lallemand (2005)
M_{max}	Maximum earthquake magnitude	–	8.226	0.6094	Heuret et al. (2011)

[a] Mean value

[b] Standard deviation

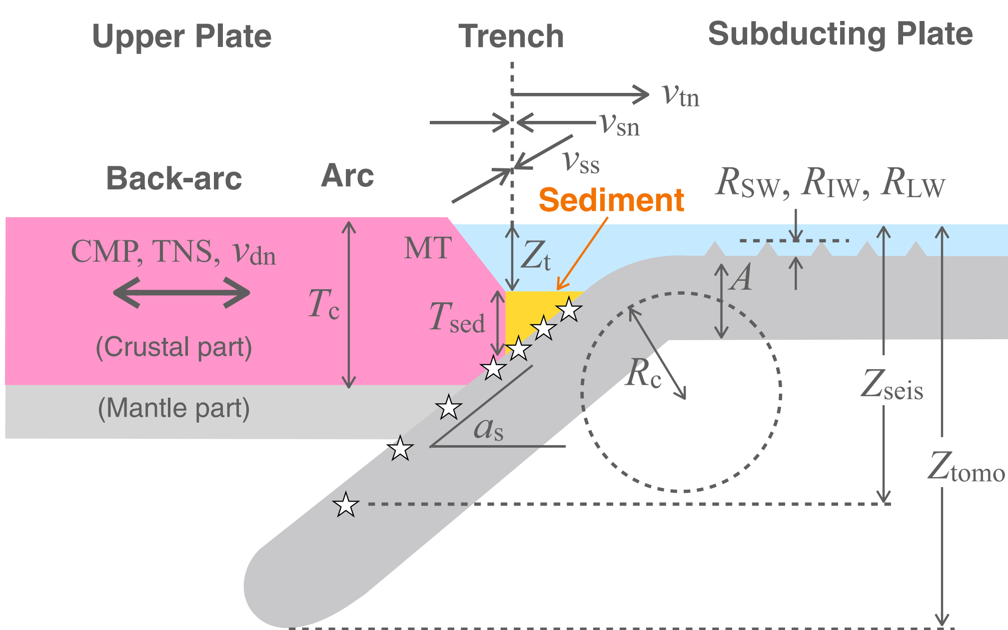


Fig. 2 Schematic cross-section of a subduction-zone, with the analyzed subduction-zone parameters labeled. The stars indicate earthquake hypocenters. The gray, pink, yellow, and blue regions indicate lithospheric rocks, the upper crust, trench sediments, and seawater, respectively. See Table 1 for details of each parameter

the subduction-zone parameters that are dependent on M_{max} by definition, such as the equivalent representative magnitude (M_{MRR}; the earthquake magnitude calculated from MRR (moment release rate), where MRR is the integrated seismic moment during a century and along 1000 km of the trench), from the regression analysis. The trench length (or slab width), which is a potential controlling factor for $M_{\max}$ (Schellart and Rawlinson 2013; Brizzi et al. 2018), is not used as an explanatory variable of $M_{\max}$ in this study, because there can be a spurious correlation between $M_{\max}$ and the trench length; more precisely, a smaller $M_{\max}$ is generally expected to be observed within a limited timeframe as the trench length becomes smaller, even if the occurrence of a large earthquake is completely random.

The details of the 18 analyzed subduction-zone variables are as follows. A is the age of the subducting oceanic plate at the trench (Müller et al. 1997). a_s is the mean dip angle of the subducting oceanic plate over the 0–125 km depth range, which is measured using hypocenters of Engdahl et al. (1998) along Wadati–Benioff zones and plate boundaries (Lallemand et al. 2005). CMP and TNS (compression and tension, respectively) are dummy parameters, which take 0 or 1, to express the strain state of the upper plate ("UPS" in Heuret et al. 2011): $(\mathrm{CMP}, \mathrm{TNS}) = (1, 0)$ for compressible upper plates, (0, 1) for tensile upper plates, and (0, 0) for neutral upper plates. The upper-plate stress is originally classified using an "ordinal scale" into three types based on the focal mechanisms of shallow earthquakes occurring at depths less than 40 km (Heuret et al. 2011); the two dummy variables CMP and TNS are necessary to express the ordinal scale in the regression modeling. MT, or the margin type, is a dummy variable that expresses either the accretionary or erosional conditions of the upper-plate margin: $\mathrm{MT} = 1$ for accretionary margins and 0 for erosional margins. R_c is the bending radius of the subducting oceanic plate, which is measured such that a circle of radius R_c within a trench-normal vertical cross section fits hypocenters of Engdahl et al. (1998) along the upper limit of the Wadati–Benioff zone over the 0–150 km depth range (Heuret 2005; Wu et al. 2008). R_{SW}, R_{IW}, and R_{LW} are the seafloor roughnesses of the subducting oceanic plate for different bathymetric wavelengths, which have been defined by Lallemand et al. (2018): 12–20 km (short wavelengths), 20–80 km (intermediate wavelengths), and 80–100 km (long wavelengths), respectively. T_c is the thickness of the margin of the upper plate, which is taken from CRUST 1.0 (Laske et al. 2013). Here, T_c is defined as the maximum thickness of the crust from the trench to the volcanic arc. T_{sed} is the sediment thickness at the trench, which is taken from GlobSed 3 (Straume et al. 2019). v_{sn} and v_{ss} are the trench-normal and -parallel components, respectively, of the convergence rate at the trench (Lallemand et al. 2008). v_{dn} is the trench-normal component of the extension rate of the upper-plate margin (Lallemand et al. 2008). v_{dn} is positive for back-arc spreading and negative for back-arc shortening. v_{tn} is the trench-normal component of the trench migration rate, with a positive value for a retreating trench (oceanward motion) and negative value for an advancing trench (continent-ward motion). We referenced the absolute velocity v_{tn} from the SB04 model (Steinberger et al. 2004; Lallemand et al. 2008), which is adjusted using the Indo-Atlantic hotspots as a reference. We referenced SB04 because the Indo-Atlantic hotspot reference frame better explains the geometry of subducting slabs beneath global subduction zones, which is sensitive to trench motion (Schellart 2011; Schellart and Rawlinson 2013; Nakao et al. 2022). We additionally applied the v_{tn} based on the Pacific hotspot reference frame (HS3; Gripp and Gordon 2002) to confirm that the influence of the reference frame on the analytical results is small, as shown in Additional file 1: Fig. S7. Z_{seis} is the maximum depth of deep earthquakes. Z_t is the trench depth. Z_{tomo} is the maximum depth of the subducting plate, which has been constrained from high-velocity seismic anomalies (Heuret and Lallemand 2005).

Regression analysis

We relate $M_{\max}$ (objective variable) to the subduction-zone parameters (explanatory variables) via regression analysis. We evaluate the contribution of each explanatory variable to $M_{\max}$ by standardizing the i-th explanatory variables at location j as follows:

$$x'_{ij} = \frac{x_{ij} - m_i}{s_i}, \tag{1}$$

where x_{ij} is an unstandardized explanatory variable (i.e., $A, a_s, \ldots, Z_{\mathrm{tomo}}$), $m_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$ is the empirical mean value of variable i, $s_i = \{\frac{1}{J-1}\sum_{j=1}^{J}(x_{ij} - m_i)^2\}^{\frac{1}{2}}$ is the empirical standard deviation of variable i, and J is the number of locations used for training. We randomly selected 95% of the 169 locations in Fig. 1 as training data for the regression analysis, with the remaining 5% used as test data to validate the optimal models (i.e., $J = 161$). Hereafter, a standardized variable is expressed using a prime symbol.

We assume that $M_{\max}$ is a linear combination of the explanatory variables:

$$f'_j(\mathbf{a};\ \mathbf{c}) = a_0 + \sum_{i=1}^{I} c_i a_i x'_{ij} + \varepsilon_j, \tag{2}$$

where $f'_j(\mathbf{a};\ \mathbf{c})$ is the predicted maximum earthquake magnitude at location j, $\mathbf{a} = (a_0, a_1, \ldots, a_I)$ is a vector of the coefficients, I is the number of explanatory variables ($I = 18$), $\mathbf{c} = (c_1, \ldots, c_I)$ is a vector of parameters that control whether x'_{ij} is included in the regression analysis (i.e., $c_i = 0$ or 1), and ε_j is Gaussian observation noise. A linear model is determined when $\mathbf{c}$ is fixed, with this linear model specified by the configuration of $\mathbf{c}$. Although such a simple linear combination of subduction-zone

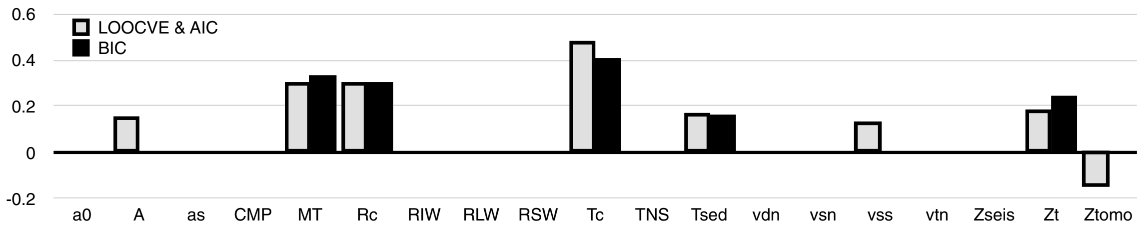


Fig. 3 Results for determining the subduction-zone parameters that are related to $M_{\max}$. The graph shows the coefficients, a_i, for each standardized explanatory variable that yields the smallest LOOCVE and AIC (gray bars; Eq. 9), and BIC (black bars; Eq. 11). An explanation of the symbols is provided in Table 1

parameters is often used to express $M_{\max}$ (e.g., Brizzi et al. 2018), this assumption is not physically derived. We therefore conduct an error analysis to check the validity of this assumption.

The coefficients are estimated via a least-squares method for a given $\mathbf{c}$, such that:

$$\mathbf{a}_{\mathrm{LS}} = \arg\min_{\mathbf{a}} \sum_{j=1}^{J} \left(f_j'(\mathbf{a}; \mathbf{c}) - M'_{\max,j} \right)^2, \tag{3}$$

where $M'_{\max,j}$ is the maximum earthquake magnitude observed at location j.

Multicollinearity, which occurs when a pair of explanatory variables has a significantly large correlation coefficient, causes problems during multiple regression analysis. For example, strong multicollinearity may prohibit the selection of an explanatory variable, even if that explanatory variable has a significant relationship to the objective variable. We verified that each of the pairs among the 18 explanatory variables have variance inflation factors (VIFs; an index for multicollinearity) below ~ 3.5 (Additional file 1: Fig. S2), thereby indicating that multicollinearity should not largely impact our results (Hair et al. 2009).

Model selection

We assume that a limited number of subduction-zone parameters essentially characterizes $M_{\max}$, such that we can determine the best combination of the subduction-zone parameters from our analysis. We select the optimal models from 2^I (=262,144) cases by calculating three criteria, the leave-one-out cross-validation error (LOOCVE), Akaike information criterion (AIC) (Akaike 1974), and Bayesian information criterion (BIC) (Schwarz 1978), for each case.

Cross-validation is a method that divides the samples into test and training data sets, with the model performance iteratively evaluated based on the data sizes. Leave-one-out cross-validation is a special case whereby one sample is left as test data to evaluate a model that is generated using the other samples, with this procedure repeated based on the number of samples:

$$\mathrm{LOOCVE}(\mathbf{c}) = \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left(f'_{-j}(\mathbf{c}) - M'_{\max,j} \right)^2}, \tag{4}$$

where $f'_{-j}(\mathbf{c})$ is the maximum earthquake magnitude at the j-th location that was predicted by the model using the other locations. Cross-validation can enhance the generalized performance of a prediction model using a small number of observations, which is common in subduction zones.

AIC and BIC are evaluation criteria that balance the misfit with the number of explanatory variables. We can avoid overfitting to noisy subduction-zone parameters by implementing either AIC or BIC. AIC and BIC are defined as

$$\mathrm{AIC}(\mathbf{c}) = -2 \ln L(\mathbf{c}) + 2 \left(\sum_{i=1}^{I} c_i + 2 \right) \tag{5}$$

and

$$\mathrm{BIC}(\mathbf{c}) = -2 \ln L(\mathbf{c}) + \left(\sum_{i=1}^{I} c_i + 2 \right) \ln J, \tag{6}$$

respectively, where $L(\mathbf{c})$ is the maximum likelihood of the model. The term $\left(\sum_{i=1}^{I} c_i + 2 \right)$ in Eqs. (5) and (6) indicates the number of parameters; the number of a_i values used in model $\mathbf{c}$, intercept a_0, and observation noise are all counted as parameters in these evaluations. $L(\mathbf{c})$ is calculated as

$$\ln L(\mathbf{c}) = -\frac{J}{2} \ln(2\pi \hat{\sigma}^2(\mathbf{c})) - \frac{J}{2}, \tag{7}$$

where $\hat{\sigma}(\mathbf{c})$ is the maximum likelihood estimate of the error variance, which can be written as

$$\hat{\sigma}^2(\mathbf{c}) = \frac{1}{J}\sum_{j=1}^{J}\left(f_j'(\mathbf{a}_{LS};\mathbf{c}) - M'_{\max,j}\right)^2. \tag{8}$$

Exhaustive model evaluation was conducted to determine the smallest LOOCVE, AIC, and BIC values, which both enhanced the predictability and minimized the overfitting of the final model.

Results

Optimal models

We obtained minimum LOOCVE and AIC values among the 2^{18} cases by characterizing $M_{\max}$ using eight explanatory variables: A, MT, R_c, T_c, T_{sed}, v_{ss}, Z_t, and Z_{tomo} (Fig. 3). The standardized and unstandardized forms of the optimal model are

$$f'_{LOOCVE} = f'_{AIC} = 5.4\times10^{-16} + 0.15A' + 0.30\mathrm{MT}' + 0.30R'_c + 0.46T'_c + 0.16T'_{sed} + 0.13v'_{ss} + 0.17Z'_t - 0.14Z'_{tomo} \tag{9}$$

and

$$f_{LOOCVE} = f_{AIC} = 6.7 + 2.1\times10^{-3}A + 0.37\mathrm{MT} + 9.1\times10^{-4}R_c + 0.018T_c + 0.13T_{sed} + 4.4\times10^{-3}v_{ss} + 0.078Z_t - 2.7\times10^{-4}Z'_{tomo}, \tag{10}$$

respectively. The minimum BIC is obtained when the model includes only five parameters, which are also used in the f'_{LOOCVE} and f'_{AIC} (Fig. 3). The standardized and unstandardized forms of the optimal model in terms of BIC are

$$f'_{BIC} = 5.53\times10^{-16} + 0.32\mathrm{MT}' + 0.30R'_c + 0.40T'_c + 0.16T'_{sed} + 0.24Z'_t \tag{11}$$

and

$$f_{BIC} = 6.7 + 0.40\mathrm{MT} + 9.0\times10^{-4}R_c + 0.015T_c + 0.12T_{sed} + 0.081Z_t, \tag{12}$$

respectively. Other parameters, including the angle of the subducting oceanic plate, seafloor roughness, upper-plate strain, and trench-normal plate velocities, were not selected as explanatory parameters in the both optimal models. Although the R_{IW}–R_{LW} pair has a slightly large VIF (3.5; Additional file 1: Fig. S2), neither is selected; therefore, our regression analysis did yield very weak multicollinearity (Hair et al. 2009).

The upper-plate crustal thickness, T_c, makes the largest contribution to each optimal model among the selected explanatory variables. In fact, all large earthquakes ($M > 9$) have occurred along trenches where the upper plate consists of continental lithosphere; e.g., the 1960 Chile M9.5, 1964 Alaska M9.2, 2004 Sumatra–Andaman M9.1, 2011 Northeast Japan 9.1, and 1952 Kamchatka M9.0 earthquakes. Conversely, no large earthquakes ($M > 8.5$) have been observed along the Mariana, Tonga–Kermadec, and South Sandwich subduction zones, where the upper plate consists of oceanic lithosphere, which has a homogeneous crustal thickness of ~ 7 km (White et al. 1992).

The margin type, MT, and the trench sediment thickness, T_{sed}, also make large positive contributions to f_{LOOCVE}, f_{AIC}, and f_{BIC}, followed by T_c. An accretionary prism (MT $= 1$) generally develops where there are thick oceanic sediments, such that both variables have a strong positive correlation (Additional file 1: Fig. S1). Accretionary prisms and large T_{sed} values are found along the Cascadia, Alaska, Antilles, Andaman, and Hikurangi subduction zones, and some great earthquakes, such as the 1964 Alaska M9.2 and 2004 Sumatra–Andaman M9.1 earthquakes, occurred in these regions.

Comparison between f_{BIC} and $M_{\max}$

Hereafter, we compare the observed $M_{\max}$ with the optimal model that yields the minimum BIC (f_{BIC}) for simplicity. This is because BIC generally yields a simpler model than LOOCVE and AIC and better fits our purpose. f_{LOOCVE} (f_{AIC}) yields approximately the same values as f_{BIC} (Additional file 1: Figs. S3, S4, S6).

Figure 4a shows that the optimal model f_{BIC} can predict $M_{\max}$ from both the test and training data sets within the 95% prediction intervals, whereas some of the f_{BIC} values along the (A) South Kermadec and (B) South-Central Chile subduction zones are outside of the prediction intervals (Fig. 4a, c). Possible reasons for the limitations of our analysis along these three subduction zones will be discussed in the Discussion section.

Figure 4b shows that the error between the predicted and maximum earthquake magnitudes, $f'_{BIC} - M'_{\max}$, possesses a Gaussian-like distribution. A Q–Q plot, which can be used to quantify the distribution of the error (Additional file 1: Fig. S5b), shows that the error $f'_{BIC} - M'_{\max}$ is well aligned with the theoretical Gaussian distribution.

Discussion

Possible effects of subduction-zone parameters on $M_{\max}$

Our analysis indicates that the trench sediment thickness, T_{sed}, is an essential factor for obtaining a large $M_{\max}$, which is consistent with previous studies (e.g., Heuret et al. 2012; Brizzi et al. 2018). There are several explanations of the effects of oceanic sediments on $M_{\max}$. One is that the subducted sediment layer creates structural coherence between two converging plates, establishing the potential for plate locking due to diagenesis (e.g., Ruff 1989). Another explanation focuses on the

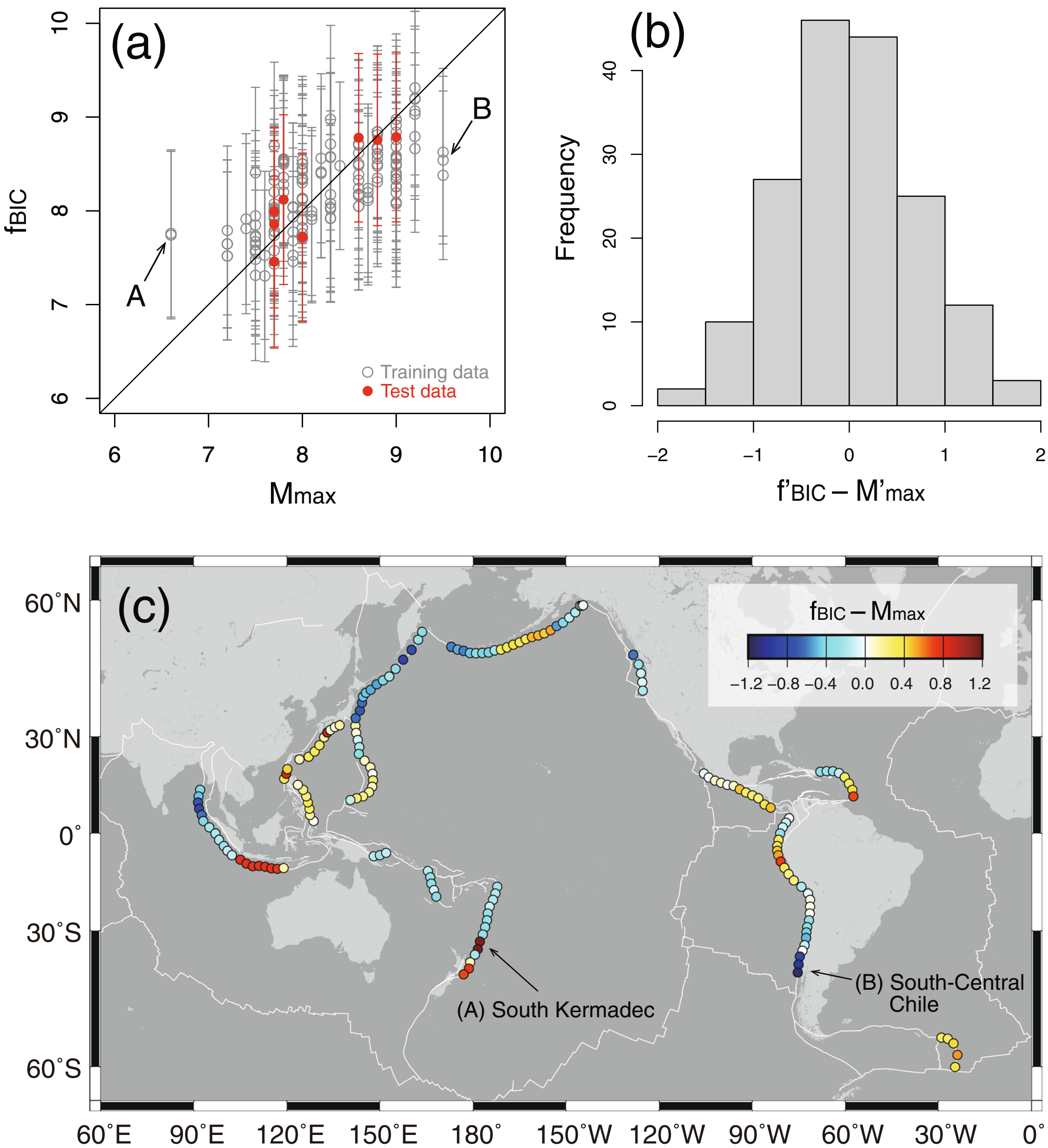


Fig. 4 Comparison between observed and modeled maximum earthquake magnitudes. **a** Unstandardized observed maximum earthquake magnitude, M_{max} (horizontal axis), versus the predicted value, f_{BIC} (vertical axis), at the 169 analyzed locations. The open gray circles and closed red circles indicate the training and test data sets, respectively. The error bars indicate the 95% prediction intervals. $f_{BIC} = M_{max}$ along the black line. Samples A (South Kermadec) and B (South-Central Chile) are outliers. **b** Histogram of the error between the predicted and observed standardized maximum earthquake magnitudes ($f'_{BIC} - M'_{max}$). **c** Map of the error between the predicted and observed unstandardized maximum earthquake magnitudes ($f_{BIC} - M_{max}$) at the 169 analyzed locations. Warm colors indicate $f_{BIC} > M_{max}$, and cool colors indicate $f_{BIC} < M_{max}$. The white lines indicate plate boundaries

small permeability of the subducted sediments (Seno 2017). The subducted oceanic crust dehydrates as the temperature and pressure increase, and the released fluid migrates toward the overlying sediment layer. Seno (2017) proposed that a thick sediment layer will act as an impermeable layer, preventing the migration of fluid from the underlying crustal layer, and the pore-fluid pressure along the subducted plate interface above the sediment layer will remain small. This mechanism may account for the positive correlation between the stress drop due to megaquakes and sediment thickness (Seno 2017).

Our study reveals that the upper-plate crustal thickness, T_c, is another essential factor in generating large M_{max}. This result is consistent with Heuret et al. (2011), who showed that continental upper plates can host larger earthquakes than oceanic upper plates. A possible explanation for the positive relationship between T_c and M_{max} is that the rupture areas (i.e., exponential function of earthquake magnitude) of large earthquakes are roughly limited to the brittle crust of the upper plate. This is because the serpentine in the mantle component of the upper plate along the plate boundary, which generally forms via plate dehydration and has a low shearing strength, tends to inhibit shear stress accumulation and instead release it via ductile deformation (e.g., Katayama et al. 2012). A key exception is the 2011 Tohoku earthquake, which occurred where the upper continental crust is slightly thinner ($\sim$ 30 km); however, the main rupture area is still estimated to be at shallow depths near the Japan Trench, with insignificant slip detected below the continental Moho (e.g., Yagi and Fukahata 2011).

The bending radius, R_c, was selected as an explanatory variable for M_{max} in our analysis, but the dip angle, a_s, was not. This result is consistent with a previous observational study (Bletery et al. 2016), in which the bending radius has the stronger correlation with M_{max} compared to the dip angle. This is probably due to the fact that a_s is the average angle for the entire slab depth (0–125 km), whereas large earthquakes generally occur at shallow depths (0–70 km). Therefore, R_c has a stronger relationship with M_{max}. We propose that, where R_c is larger, more of the plate boundary is in contact with the crustal part of the upper plate, along which elastic stress would accumulate.

It is unclear why the trench depth, Z_t, is selected in the optimal model and why its coefficient is positive, particularly since Z_t generally reflects negative slab buoyancy and is expected to have a negative contribution to the earthquake size (Nishikawa and Ide 2014). Although Z_t is negatively correlated with T_{sed} as sediment fills a trench, multicollinearity does not greatly affect our results due to the small correlation coefficient (~ -0.4; Additional file 1: Figs. S1 and S2), and Z_t would have a role independent of T_{sed}. An explanation for the positive relationship between Z_t and M_{max} is that, when a subductinhg slab contains a large amount of water, both of Z_t and M_{max} become small. That is, subducting plates containing a large amount of water have a large positive buoyancy (Nakao et al. 2016, 2018), thereby yielding small Z_t values. Meanwhile, such a plate would cause significant dehydration and subsequently yield large pore-fluid pressures along the plate boundary, inhibiting stress accumulation along the plate boundary. However, this explanation is highly speculative; numerical simulations are required to reveal the physical mechanisms that may induce large earthquakes along deep trenches.

A smooth seafloor is often regarded as an essential factor in generating large earthquakes (e.g., Wang and Bilek 2014) because a smooth plate interface contributes to a coherent plate boundary. However, the seafloor roughness, at least at wavelengths greater than 12 km, is not selected as an explanatory variable in our analysis. A possible reason for this omission is that the subducted sediment layers are up to ~1 .6 km thick (Seno 2017), which may cover and smooth a large percentage of the seafloor relief, whereas the typical seafloor roughness is $\sim$ 0.5 km, with a standard deviation of $\sim$ 0.5 km (Table 1). This may be the reason why the seafloor roughness is not selected as a universal explanatory variable in our analysis.

The upper-plate stress was not selected as an explanatory variable of M_{max} in our analysis, even though it had been selected in some previous studies (Heuret et al. 2011, 2012). Our results suggest that the present-day stress is not a good indicator for evaluating the potential M_{max}. A possible interpretation of our result is that the stress patterns change temporally throughout earthquake cycles. For example, the upper-plate stress immediately changed from compressional to tensional, as observed before and after the 2011 Tohoku earthquake (Hasegawa et al. 2012).

Thus, we propose that multiple factors influence M_{max}. For example, the 1868 Peru earthquake (Mw8.5–9.2; Lay and Nishenko 2022; McCaffrey 2008) has occurred where the trench sediment is thin (< 1 km), suggesting that it is difficult to attribute the magnitude of this earthquake to the sole role of the sediment. Rather, this historical earthquake may be related to the thick continental crust ($T_c \sim 60$ km), which is newly found as a factor for large M_{max} in our study. However, it is difficult to generate a $M \geq 8$ earthquake considering a typical value of T_c; therefore, the combined effects of other factors, including T_{sed} and R_c, are necessary to generate a $M \geq 8$ earthquake.

Previous studies have proposed numerous possible mechanisms for the genesis of large earthquakes as in Introduction. Our exhaustive model evaluation has detected the subduction-zone parameters that possess a strong relationship with M_{max}, which enables us to evaluate the plausibility of the proposed mechanisms for generating large earthquakes. Our presented approach will therefore assist in clarifying problems that include complex processes.

Origin of the misfit between f_{BIC} and M_{max}

Here we discuss why there is high degree of misfit between f_{BIC} and M_{max} at some of the analyzed locations (Fig. 4a). There are large f_{BIC} values along the South

Kermadec subduction zone (A in Fig. 4a, c) because the subduction-zone parameters are almost the same as those along the North Kermadec subduction zone, whereas $M_{\max}$ is significantly smaller. If our analysis accurately reflects a sufficient number of factors for constraining the earthquake magnitude, then a M-8 class earthquake will potentially occur along the South Kermadec subduction zone. Or, the North and South Kermadec regions should be integrated into the same group. The South-Central Chile subduction zone (B in Fig. 4a, c) is another outlier. These results suggest that the 1960 Chile earthquakes are linked to tectonic processes that are not captured by the subduction parameters considered in our analysis, which focuses on large-scale tectonic features. Ridge subduction, petit-spots, and hydrothermal circulation are potential candidates for the elevated earthquake magnitudes along the South-Central Chile subduction zone. However, we cannot identify the factor(s) here, but hope to resolve this in a future study.

The misfit between the predicted and observed maximum earthquake magnitudes, $f_{\mathrm{BIC}} - M_{\max}$, yields a Gaussian-like distribution among the 169 analyzed locations. The least-squares method, which is used in our regression analysis, is based on the assumption that the error follows a Gaussian distribution. Therefore, it is suggested that our modeling, which is based on the assumption that f_{BIC} is a linear combination of the subduction-zone parameters, is not unreasonable. In addition, the Gaussian-like distribution of the misfit justifies applying $M_{\max}$ in Fig. 1 as the objective variable to some extent although $M_{\max}$ may lack some unknown historical earthquakes because of the observation duration shorter than megathrust earthquake cycles (10^2–10^3 years). This would be because large earthquakes have been observed in this 10^2 year correspondingly to potential (or ideal) $M_{\max}$ following the Gutenberg–Richter's law.

Conclusions

We conducted multiple regression analyses and exhaustive model evaluation to determine the subduction-zone parameters that control maximum earthquake magnitude, $M_{\max}$. The smallest LOOCVE and AIC evaluation criteria were obtained when eight parameters, the trench sediment thickness, T_{sed}, existence of an accretionary prism, MT, upper-plate thickness, T_{c}, bending radius of the subducting oceanic plate, R_{c}, trench depth, Z_{t}, age of the subducting plate, A, along-strike convergence rate along the trench, v_{ss}, and maximum depth of the subducting plate, Z_{tomo}, were selected as explanatory variables to express $M_{\max}$. Furthermore, the combination of only five variables, T_{sed}, MT, T_{c}, R_{c}, and Z_{t}, yields the smallest BIC. The seafloor roughness, trench-normal plate and trench velocities, upper-plate stress, and dip angle of the subducting oceanic plate are notable subduction-zone parameters that were not selected as explanatory variables. Our results are consistent with previous studies that have proposed T_{sed} as the primary factor controlling $M_{\max}$. We provided new insight that T_{c} also has a positive effect on producing large $M_{\max}$, which suggests that along-dip extension of crustal areas along a converging plate boundary are important in generating large earthquakes. We used five tectonic conditions, T_{sed}, MT, T_{c}, R_{c}, and Z_{t}, to demonstrate that our optimal model can explain almost all of the observed $M_{\max}$ values within the 95% confident interval, although our model fails to predict some samples, such as the 1960 M9.5 Chile earthquake. An evaluation of additional mechanisms that may cause these outliers will be conducted to understand the processes controlling the earthquakes that are not explained by our model. An investigation of the genesis of large earthquakes using numerical simulations that consider the essential subduction-zone factors for generating large $M_{\max}$ will also be undertaken, as our analyses do not explain the physical meanings of the detected parameters.

Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
LOOCVE	Leave-one-out cross-validation error
$M_{\max}$	Maximum earthquake magnitude
MT	Margine type
CMP	Compression
TNS	Tension

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40623-023-01839-y.

Additional file 1. Additional mechanisms.

Acknowledgements
We thank two anonymous reviewers for their constructive comments and Naoki Uchida and Aitaro Kato for editing this article. We used the R language (R Core Team 2022) and the `MuMIn` package (Barton and Barton 2015) for model selection. Some of the figures were produced using Generic Mapping Tools (Wessel et al. 2019).

Author contributions
AN designed the research, collected the data, conducted the analysis, and wrote the original draft under the supervision of TK. TK and KU developed the analytical method. TK, KU, KY, and TY reviewed and edited the original draft. HH and SA verified the analytical method and reviewed and edited the statistical discussions in the manuscript.

Authors' information
AN (Corresponding author, Post-doctoral Researcher), TK (Deputy Group Leader, Senior Researcher), KU (Researcher), KY (Researcher), and TY (Research Assistant) belongs to Solid-Earth Data Science Group (SDG) of JAMSTEC. TK employed AN and TY under the CREST project until March 2023, and AN is currently an assistant professor at Akita University. TK and HH (Professor of

ISM) are principal researchers of the CREST project. HH and SA (Chief Senior Researcher of AIST) work with SDG through Joint Research program of ERI.

Funding
This work was supported by Japan Science and Technology Agency CREST Grant No. JPMJCR1761; JSPS KAKENHI Grant Nos. JP19K04026, JP20K04119, JP22K14131, and JP22H03653; and the Joint Research Programs 2021-B-01 and 2022-B-06 of the Earthquake Research Institute, the University of Tokyo.

Availability of data and materials
All of the data used in this study are available from SubMap 4.3 (Heuret and Lallemand 2005, http://submap.gm.univ-montp2.fr), CRUST 1.0 (Laske et al. 2013, https://igppweb.ucsd.edu/~gabi/rem.html), and GlobSed 3 (Straume et al. 2019, https://ngdc.noaa.gov/mgg/sedthick/).

Declarations

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential competing interests.

Author details
[1]Research Institute for Marine Geodynamics, Japan Agency for Marine-Earth Science and Technology, Yokosuka 237-0061, Japan. [2]The Institute of Statistical Mathematics, Tachikawa 190-8562, Japan. [3]National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8568, Japan. [4]Graduate School of Engineering Science, Akita University, Akita 010-8502, Japan.

Received: 12 October 2022 Accepted: 3 May 2023
Published online: 12 May 2023

References
Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
Bletery Q, Thomas AM, Rempel AW, Karlstrom L, Sladen A, De Barros L (2016) Mega-earthquakes rupture flat megathrusts. Science 354(6315):1027–1031
Brizzi S, Sandri L, Funiciello F, Corbi F, Piromallo C, Heuret A (2018) Multivariate statistical analysis to investigate the subduction zone parameters favoring the occurrence of giant megathrust earthquakes. Tectonophysics 728:92–103
Engdahl ER, van der Hilst R, Buland R (1998) Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. Bull Seismol Soc Am 88(3):722–743
Fujiwara O, Aoshima A, Irizuki T, Ono E, Obrochta SP, Sampei Y, Sato Y, Takahashi A (2020) Tsunami deposits refine great earthquake rupture extent and recurrence over the past 1300 years along the Nankai and Tokai fault segments of the Nankai Trough, Japan. Quat Sci Rev 227:105999
Gripp AE, Gordon RG (2002) Young tracks of hotspots and current plate velocities. Geophys J Int 150(2):321–361
Hasegawa A, Yoshida K, Asano Y, Okada T, Iinuma T, Ito Y (2012) Change in stress field after the 2011 great Tohoku–Oki earthquake. Earth Planet Sci Lett 355:231–243. https://doi.org/10.1016/j.epsl.2012.08.042
Heuret A, Lallemand S (2005) Plate motions, slab dynamics and back-arc deformation. Phys Earth Planet Inter 149(1–2):31–51. https://doi.org/10.1016/j.pepi.2004.08.022
Heuret A, Lallemand S, Funiciello F, Piromallo C, Faccenna C (2011) Physical characteristics of subduction interface type seismogenic zones revisited. Geochem Geophys Geosyst 12(1):Q01004
Heuret A, Conrad CP, Funiciello F, Lallemand S, Sandri L (2012) Relation between subduction megathrust earthquakes, trench sediment thickness and upper plate strain. Geophys Res Lett 39(5):L05304
Igarashi Y, Takenaka H, Nakanishi-Ohno Y, Uemura M, Ikeda S, Okada M (2018) Exhaustive search for sparse variable selection in linear regression. J Phys Soc Jpn 87(4):044802
Itano K, Ueki K, Iizuka T, Kuwatani T (2020) Geochemical discrimination of monazite source rock based on machine learning techniques and multinomial logistic regression analysis. Geosciences 10(2):63
Katayama I, Terada T, Okazaki K, Tanikawa W (2012) Episodic tremor and slow slip potentially linked to permeability contrasts at the Moho. Nat Geosci 5(10):731–734
Kuwatani T, Nagata K, Okada M, Watanabe T, Ogawa Y, Komai T, Tsuchiya N (2014) Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. Sci Rep 4(1):7077
Lallemand S, Heuret A, Boutelier D (2005) On the relationships between slab dip, back-arc stress, upper plate absolute motion, and crustal nature in subduction zones. Geochem Geophys Geosyst 6(9):Q09006
Lallemand S, Heuret A, Faccenna C, Funiciello F (2008) Subduction dynamics as revealed by trench migration. Tectonics 27(3):TC3014
Lallemand S, Peyret M, van Rijsingen E, Arcay D, Heuret A (2018) Roughness characteristics of oceanic seafloor prior to subduction in relation to the seismogenic potential of subduction zones. Geochem Geophys Geosyst 19(7):2121–2146
Laske G, Masters G, Ma Z, Pasyanos M (2013) Update on CRUST1.0-A 1-degree global model of Earth's crust. Geophys Res Abstr 15(15):2658
Lay T, Nishenko SP (2022) Updated concepts of seismic gaps and asperities to assess great earthquake hazard along South America. Proc Natl Acad Sci 119(51):e2216843119
Marzocchi W, Sandri L, Heuret A, Funiciello F (2016) Where giant earthquakes may come. J Geophys Res Solid Earth 121(10):7322–7336
McCaffrey R (2008) Global frequency of magnitude 9 earthquakes. Geology 36(3):263–266
Müller RD, Roest WR, Royer JY, Gahagan LM, Sclater JG (1997) Digital isochrons of the world's ocean floor. J Geophys Res Solid Earth 102(B2):3211–3214
Nakao A, Iwamori H, Nakakuki T (2016) Effects of water transportation on subduction dynamics: roles of viscosity and density reduction. Earth Planet Sci Lett 454:178–191. https://doi.org/10.1016/j.epsl.2016.08.016
Nakao A, Iwamori H, Nakakuki T, Suzuki YJ, Nakamura H (2018) Roles of hydrous lithospheric mantle in deep water transportation and subduction dynamics. Geophys Res Lett 45(11):5336–5343
Nakao A, Kuwatani T, Ueki K, Yoshida K, Yutani T, Hino H, Akaho S (2022) Subduction-zone parameters that control slab behavior at the 660-km discontinuity revealed by logistic regression analysis and model selection. Front Earth Sci 10:1008058
Nishikawa T, Ide S (2014) Earthquake size distribution in subduction zones linked to slab buoyancy. Nat Geosci 7(12):904–908
R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Ruff LJ (1989) Do trench sediments affect great earthquake occurrence in subduction zones? Pure Appl Geophys 129:263–282
Ruff L, Kanamori H (1980) Seismicity and the subduction process. Phys Earth Planet Inter 23(3):240–252. https://doi.org/10.1016/0031-9201(80)90117-X
Satake K, Shimazaki K, Tsuji Y, Ueda K (1996) Time and size of a giant earthquake in Cascadia inferred from Japanese tsunami records of January 1700. Nature 379(6562):246–249
Schellart WP (2011) A subduction zone reference frame based on slab geometry and subduction partitioning of plate motion and trench migration. Geophys Res Lett 38(16):L16317
Schellart WP, Rawlinson N (2013) Global correlations between maximum magnitudes of subduction zone interface thrust earthquakes and physical parameters of subduction zones. Phys Earth Planet Inter 225:41–67. https://doi.org/10.1016/j.pepi.2013.10.001
Scholl DW, Kirby SH, von Huene R, Ryan H, Wells RE, Geist EL (2015) Great ($Mw \geq 8.0$) megathrust earthquakes and the subduction of excess sediment and bathymetrically smooth seafloor. Geosphere 11(2):236–265
Scholz CH, Campos J (1995) On the mechanism of seismic decoupling and back arc spreading at subduction zones. J Geophys Res Solid Earth 100(B11):22103–22115
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Seno T (2017) Subducted sediment thickness and Mw 9 earthquakes. J Geophys Res Solid Earth 122(1):470–491
Song TRA, Simons M (2003) Large trench-parallel gravity variations predict seismogenic behavior in subduction zones. Science 301(5633):630–633

Steinberger B, Sutherland R, O'connell RJ (2004) Prediction of Emperor–Hawaii seamount locations from a revised model of global plate motion and mantle flow. Nature 430(6996):167–173
Straume EO, Gaina C, Medvedev S, Hochmuth K, Gohl K, Whittaker JM, Abdul Fattah R, Doornenbal JC, Hopper JR (2019) GlobSed: updated total sediment thickness in the world's oceans. Geochem Geophys Geosyst 20(4):1756–1772
Ueki K, Kuwatani T, Okamoto A, Akaho S, Iwamori H (2020) Thermodynamic modeling of hydrous-melt–olivine equilibrium using exhaustive variable selection. Phys Earth Planet Inter 300:106430. https://doi.org/10.1016/j.pepi.2020.106430
Wang K, Bilek SL (2014) Fault creep caused by subduction of rough seafloor relief. Tectonophysics 610:1–24
Wells RE, Blakely RJ, Sugiyama Y, Scholl DW, Dinterman PA (2003) Basin-centered asperities in great subduction zone earthquakes: A link between slip, subsidence, and subduction erosion? J Geophys Res Solid Earth 108(B10):2507
Wessel P, Luis JF, Uieda L, Scharroo R, Wobbe F, Smith WH, Tian D (2019) The generic mapping tools version 6. Geochem Geophys Geosyst 20(11):5556–5564
White RS, McKenzie D, O'Nions RK (1992) Oceanic crustal thickness from seismic measurements and rare earth element inversions. J Geophys Res Solid Earth 97(B13):19683–19715
Wirth EA, Sahakian VJ, Wallace LM, Melnick D (2022) The occurrence and hazards of great subduction zone earthquakes. Nat Rev Earth Environ 3(2):125–140
Wu B, Conrad CP, Heuret A, Lithgow-Bertelloni C, Lallemand S (2008) Reconciling strong slab pull and weak plate bending: the plate motion constraint on the strength of mantle slabs. Earth Planet Sci Lett 272(1–2):412–421. https://doi.org/10.1016/j.epsl.2008.05.009
Yagi Y, Fukahata Y (2011) Rupture process of the 2011 Tohoku–Oki earthquake and absolute elastic strain release. Geophys Res Lett 38(19):L19307
Zachariasen J, Sieh K, Taylor FW, Edwards RL, Hantoro WS (1999) Submergence and uplift associated with the giant 1833 Sumatran subduction earthquake: evidence from coral microatolls. J Geophys Res Solid Earth 104(B1):895–919
Barton K, Barton MK (2015) Package 'MuMIn'. Version 1
Hair JF, Black WC, Babin BJ, Anderson RE (2009) Multivariate data analysis. 7th Edition, Pearson
Heuret A (2005) Dynamique des zones de subduction: Etude statistique globale et approche analogique. PhD thesis, Université Montpellier II

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.